A deep learning approach for the classification of supernovae and the estimation of photometric redshifts

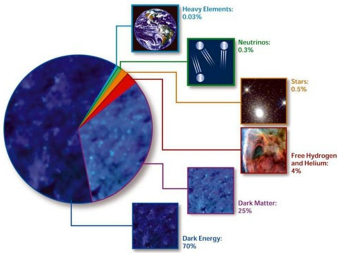
Johanna Pasquet

Centre de Physique des Particules de Marseille

Dark energy colloque

25 October, 2018

The dark energy problem



Credit : NASA

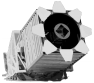
- What is the nature of dark energy ?
- Is it "dark energy" arising from quantum fluctuations in the vacuum, or is it new gravitational physics ?

The era of large surveys

1924 Henry Drapper Catalog (0.2 Million)

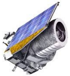
1989 Guide Star Catalog (20 Million)

2008 **SDSS (230 Million)**

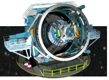


2018 **Dark Energy Survey (400 Million)**

2027 **Euclid (10 billion)**



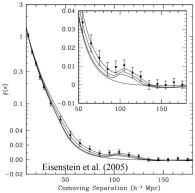
2032 **Large Synoptic Survey Telescope (37 billion)**



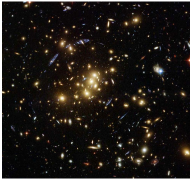
Need accurate redshifts for cosmology

Reliable redshifts are necessary to constrain the dark energy equation-of-state and to study the large scale structure of the universe

- Baryonic Acoustic Oscillations



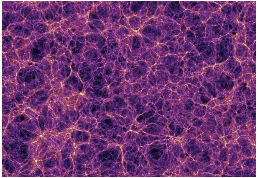
- Weak lensing



Strong gravitational lensing around galaxy cluster CL0024+17

Credit : NASA/ESA/M.J. Jee (John Hopkins University)

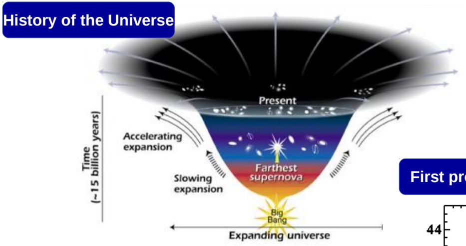
- Cosmic web

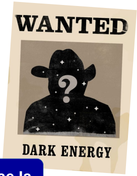


Results of a digital simulation showing the large-scale distribution of matter, with filaments and knots.

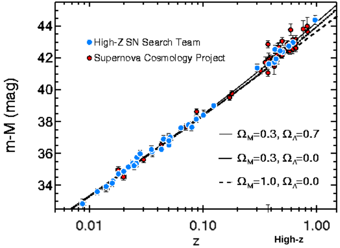
Credit: V.Springel, Max-Planck Institut für Astrophysik, Garching bei München

Supernovae Ia as cosmological probe





First proof with supernovae Ia



- Dark energy causes the universal expansion to accelerate
- Recent observations of supernovae have produced a value for an acceleration that implies a universe that is about 70 % dark energy

First application: The estimation of photometric redshift with a deep architecture

J. Pasquet, E. Bertin, M. Treyer, S. Arnouts and D. Fouchez

Photometric redshifts with Deep Learning

Photometric redshifts from SDSS images using a Convolutional Neural Network (**J. Pasquet**, E. Bertin, M. Treyer, S. Arnouts and D. Fouchez)

arxiv: 1806.06607, **code available at**: https://github.com/jpasquet/Photoz

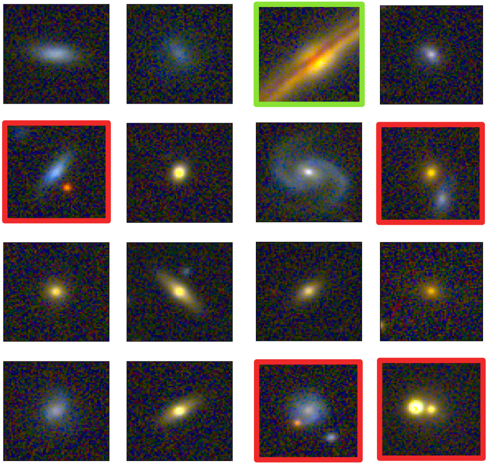
Key elements :

1. A representative and a complete training database with r-band magnitude ≤ 17.8 and redshift, $z \leq 0.4$ (516,525 galaxies)
2. Photoz values + associated Probability Distribution Functions
3. Photoz immune to IQ variations and neighbours contamination
4. A dedicated Neural Network architecture

Results obtained :

Clear improvements compared to other methods!

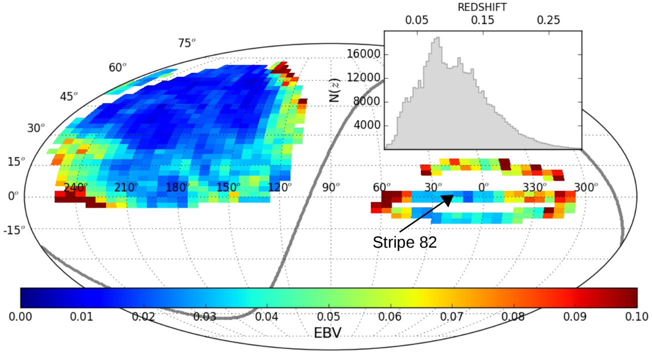
Input SDSS galaxy images transmitted to the CNN



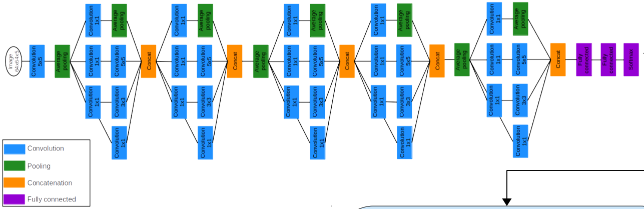
– large galaxies — crowded images

Main Galaxy Sample SDSS

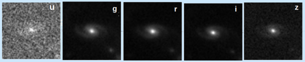
A multi-band imaging and spectroscopic redshift survey



Our architecture

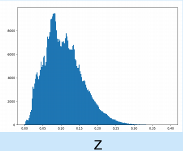


Input images of size 64x64



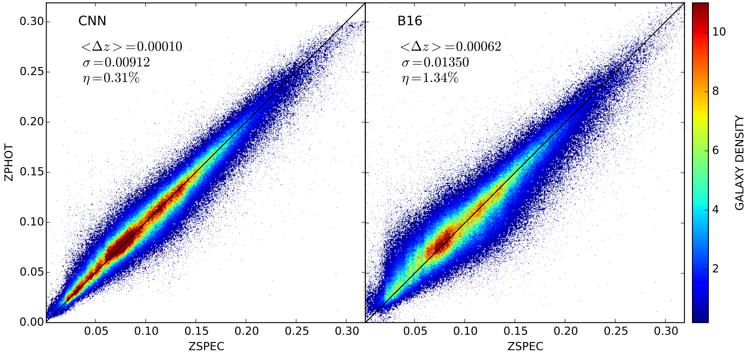
Output probabilities

Vector of posterior probabilities for the galaxy redshift to be in a z-bin



The estimator for z-phot is the centroid of the zPDF

Results of the method



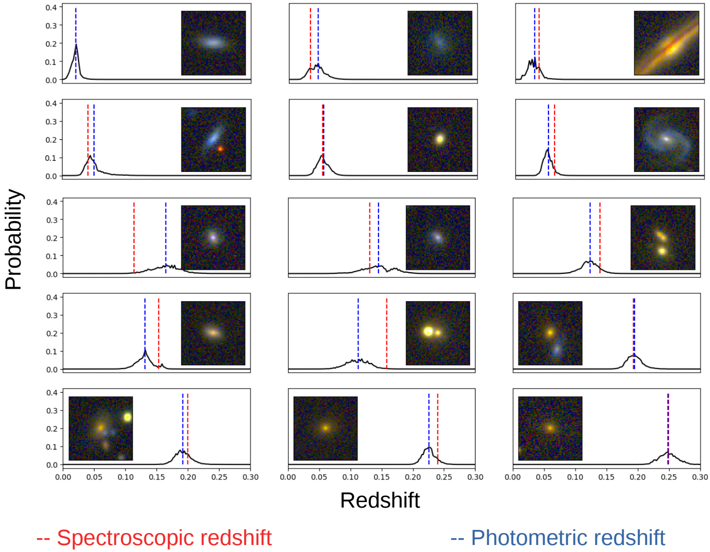
$\langle \Delta z \rangle = 1.0 \times 10^{-4}$ ← Factor of 6 improvement $\langle \Delta z \rangle = 6 \times 10^{-4}$

$\sigma = 9.1 \times 10^{-3}$ ← 30 % improvement $\sigma = 1.3 \times 10^{-2}$

$\eta = 0.31\,\%$ ← Factor of 4 improvement $\eta = 1.35\,\%$

$$\Delta z = (z_{phot} - z_{spec}) / (1 + z_{spec})$$
$$\sigma = 1.4826 \times MAD$$
$$MAD = Median(|\Delta z - Median(\Delta z)|)$$
$$\eta = |\Delta z| > 0.05$$

Examples of PDFs



Summary results

Trial	training sample size	bias	σ	η
Training with 80% of the dataset	393,219			
Full test sample		**0.00010**	**0.00912**	**0.31**
(B16)		(0.00062)	(0.01350)	(1.34)
Widest 20% of PDFs		**0.00005**	**0.00789**	**0.06**
Stripe 82 only		-0.00009	0.00727	0.34
Stripe 82 with widest 20% of PDFs removed		**0.00004**	**0.00635**	**0.09**
Training with 50% of the dataset*	250,000	0.00007	0.00910	0.29
Training with 20% of the dataset	**99,001**	**-0.00001**	**0.00914**	0.30
Training with 2% of the dataset	10,100	-0.00017	0.01433	1.26
Training and testing on Stripe 82	**15,771**	-0.00002	**0.00795**	0.38

Second application: The classification of light curves of supernovae (SN Ia/ SN Non-Ia)

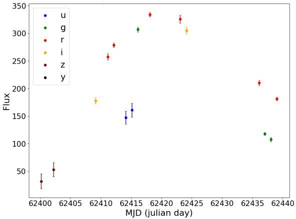
Johanna Pasquet, Jérôme Pasquet, Marc Chaumont and Dominique Fouchez

Difficulties for the classification

Many factors degrade the performance of machine learning algorithms:

Small training databases

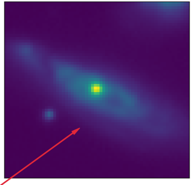
Data can be sparse with an irregular sampling



Non-representativeness between the training and the test databases

The spectroscopic follow-up

Identify and measure the redshift of a galaxy



galaxy

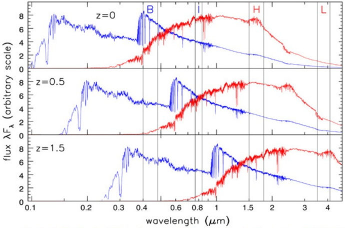
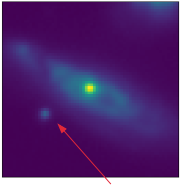
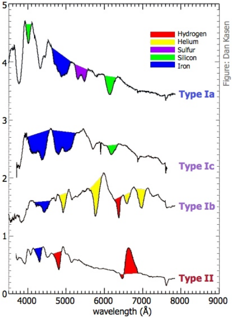


Fig 8.12 (S. Charlot) 'Galaxies in the Universe' Sparke/Gallagher CUP 2007

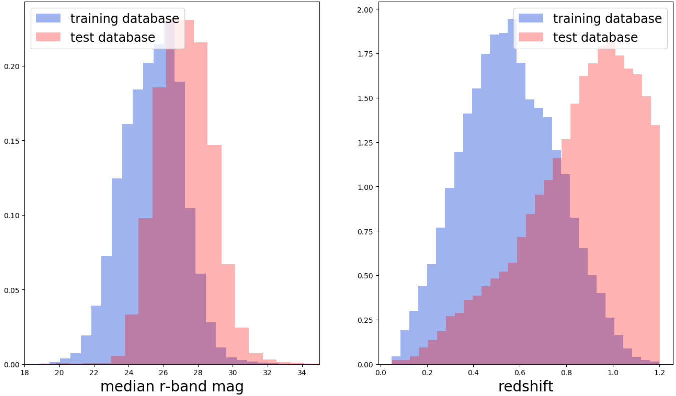
Determine the nature of an observed object



Supernovae

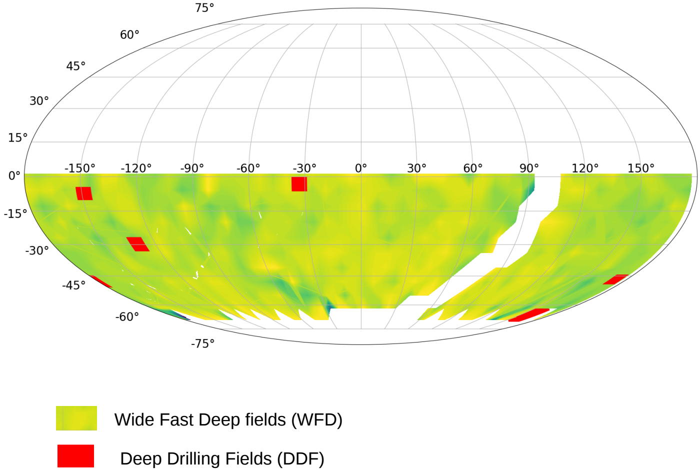


Non-representativeness between the training and test databases



The non-representativeness of the databases, which is a problem of mismatch, is critical for machine learning process.

The main survey and the deep fields of LSST

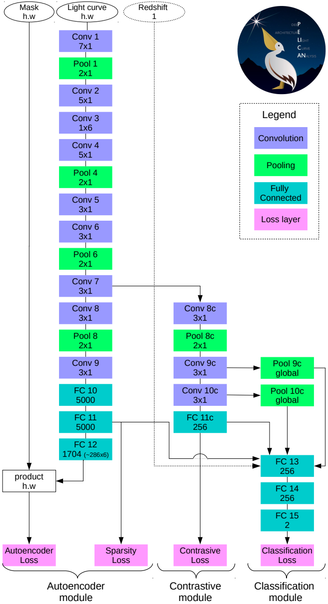


PELICAN: a deeP architecturE for the LIght Curve ANalysis
(**Johanna Pasquet**, Jérôme Pasquet, Marc Chaumont and Dominique Fouchez, just submitted)

Key elements :

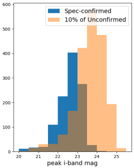
1. a complex Deep Learning architecture to classify light curves of supernovae
2. trained on a small and biased training database
3. overcome the problem of non-representativeness between the training and the test databases
4. deal with the sparsity of data and the difference of sampling and noise

The ability of PELICAN to deal with the different causes of non-representativeness between the training and test databases, and its robustness against survey properties and observational conditions, put it on the forefront of the light curves classification tools for the LSST era.


Mask
h.w
Light curve
h.w
Redshift
1
Conv 1
7x1
Pool 1
2x1
Conv 2
5x1
Conv 3
1x6
Conv 4
5x1
Pool 4
2x1
Conv 5
3x1
Conv 6
3x1
Pool 6
2x1
Conv 7
3x1
Conv 8
3x1
Pool 8
2x1
Conv 9
3x1
FC 10
5000
FC 11
5000
FC 12
1704 (~286x6)
product
h.w
Conv 8c
3x1
Pool 8c
2x1
Conv 9c
3x1
Conv 10c
3x1
FC 11c
256
Pool 9c
global
Pool 10c
global
FC 13
256
FC 14
256
FC 15
2
Legend
Convolution
Pooling
Fully Connected
Loss layer
Autoencoder Loss
Sparsity Loss
Contrasive Loss
Classification Loss
Autoencoder module
Contrastive module
Classification module

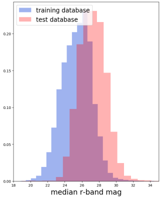
Different databases

1. The Supernova Photometric Classification Challenge in 2010 (SPCC, Kessler et al.)



- Small training database (1,103 light curves)
- Non-representativeness between the training and the test databases due to the limitation of the spectroscopic follow-up

2. LSST simulated data

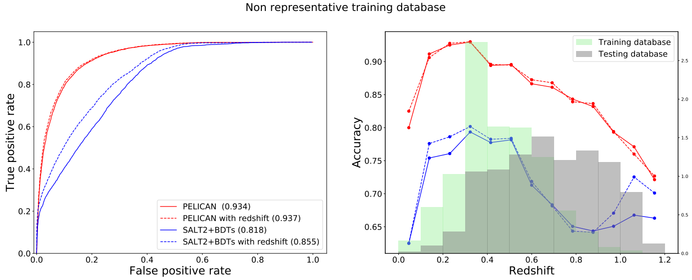


- Small training database (until 500 light curves)
- Non-representativeness between the training and the test databases due to the limitation of the spectroscopic follow-up
- Non-representativeness of the sampling and noise between main survey and deep fields

3. SDSS-II Supernova Survey Data (Frieman et al. 2008; Sako et al. 2008)

- Non-representativeness between the training (simulated data) and the test databases (real data)

The SPCC challenge

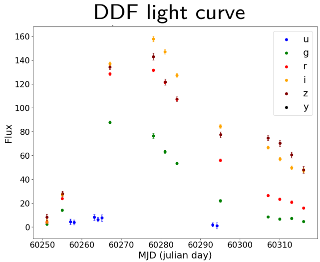


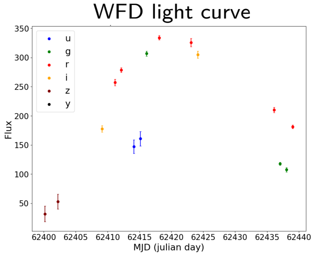
- We compared our results to BDTs classifier + SALT2 features as it is the best combination in Lochner et al. (2016)
- PELICAN obtains an accuracy of 0.856 and an AUC of 0.934 which outperforms BDTs+SALT2 method which reaches 0.705 and 0.818

LSST simulated data

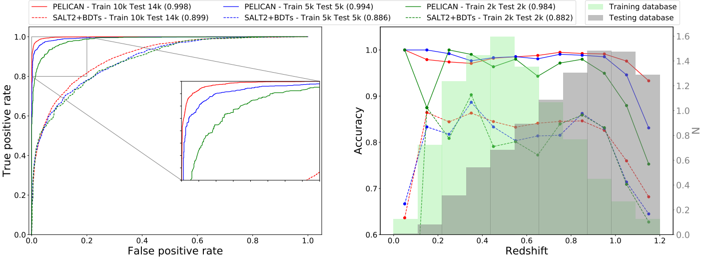
Two methodologies:

1. A training and a test on deep fields (DDF)
2. A training on deep fields and a test on the main survey (WFD)



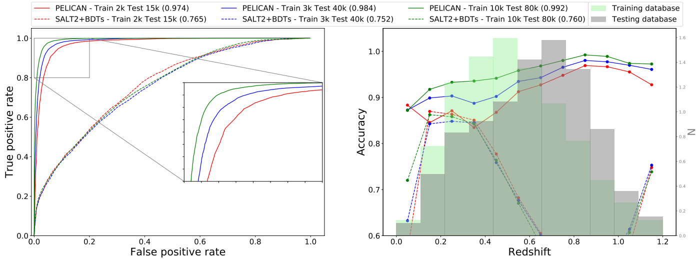


Results on DDF



	Training database (spec only)	Test database (phot only)	Accuracy	$\text{Recall}_{\text{Ia}}$ $\text{Precision}_{\text{Ia}} >$ 0.95	$\text{Recall}_{\text{Ia}}$ $\text{Precision}_{\text{Ia}} >$ 0.98	AUC
DDF	500	1,500	0.849 (0.746)	0.617 (0.309)	0.479 (0.162)	0.937 (0.848)
	2,000	2,000	0.925 (0.783)	0.895 (0.482)	0.818 (0.299)	0.984 (0.882)
	2,000	**22,000**	**0.934 (0.793)**	**0.926 (0.436)**	**0.851 (0.187)**	**0.986 (0.880)**
	10,000	14,000	0.979 (0.888)	0.992 (0.456)	0.978 (0.261)	0.998 (0.899)

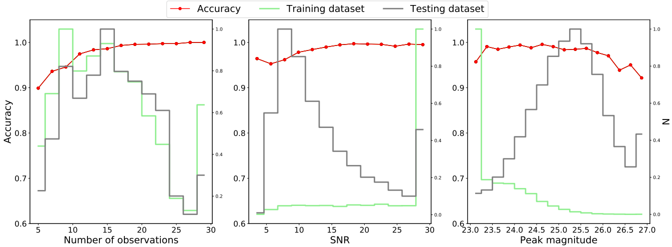
Results on WFD



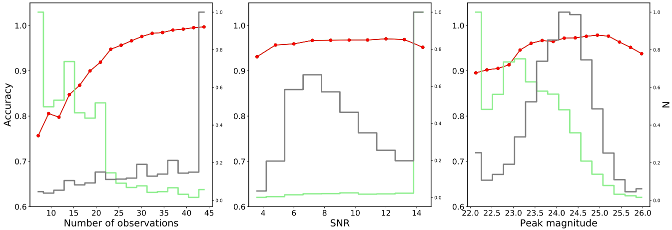
	Training database (spec only)	Test database (phot only)	Accuracy	$\mathrm{Recall}_{\mathrm{Ia}}$ $\mathrm{Precision}_{\mathrm{Ia}} > 0.95$	$\mathrm{Recall}_{\mathrm{Ia}}$ $\mathrm{Precision}_{\mathrm{Ia}} > 0.98$	AUC
WFD	DDF Spec : 2, 000	WFD : 15, 000	0.917 (0.650)	0.857 (0.066)	0.485 (0.000)	0.974 (0.765)
	DDF Spec : 3, 000	**WFD : 40, 000**	**0.940 (0.650)**	**0.939 (0.111)**	**0.729 (0.000)**	**0.984 (0.752)**
	DDF Spec : 10, 000	WFD : 80, 000	0.962 (0.651)	0.977 (0.121)	0.889 (0.010)	0.992 (0.760)

Further analysis of the behaviour of PELICAN

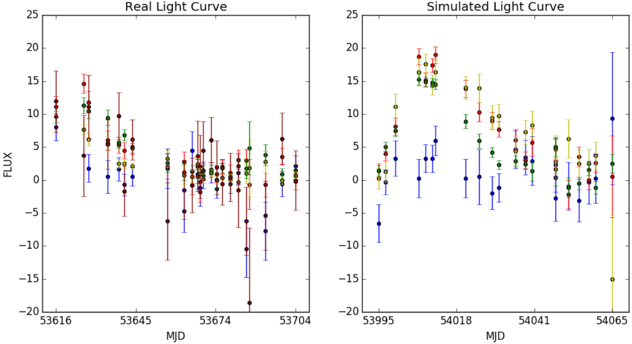
DDF



WFD



SDSS data



Training database	test database	Accuracy	AUC
SDSS simulations : 219,362	SDSS-II SN confirmed : 582	0.462	0.722
SDSS simulations : 219,362 SDSS-II SN confirmed : 80	SDSS-II SN confirmed : 582	0.868	0.850

Summary

Era of Big data

The future surveys will deliver multi-band photometry for billions of sources

Many issues for the classification algorithms

- Small size of the training database due to the limitation of the spectroscopic follow-up
- Several problems of representativeness
- Nature of data : sparse with an irregular sampling

Promising results for the estimation of photometric redshifts

We developed a CNN used as a classifier to estimate photometric redshifts and their associated PDFs. • Our work shows significant significant improvements for:

- the dispersion of photometric redshifts,
- the PDFs that are well calibrated
- no measurable bias with the reddening and the inclination of galaxies

Summary

New solutions for the classification of light curves

PELICAN obtained the best performance ever achieved with a non-representative training database of the SPCC challenge

PELICAN is able to significantly remove several types of non-representativeness between the training and the test databases due to :

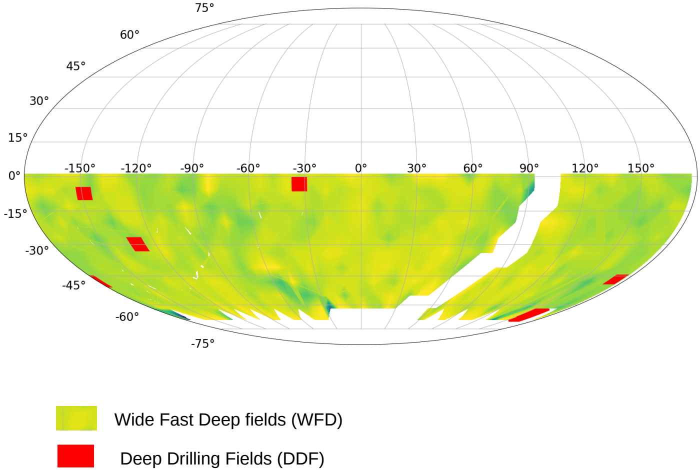
- the limit in brightness and redshift of the spectroscopically confirmed data
- the different observational strategies
- the difficulty of simulated data to reproduce perfectly real data

PELICAN can deal with the data that are sparse, with an irregular sampling

Perspectives

PELICAN offers promising perspectives for the classification of light curves and the estimation of photometric redshifts, as the method can be applied to images.

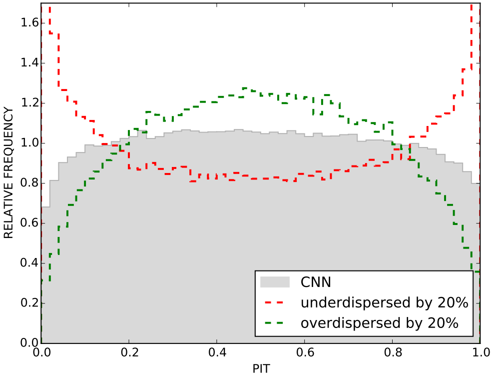
The main survey and the deep fields of LSST



Assess the prediction quality of our PDFs

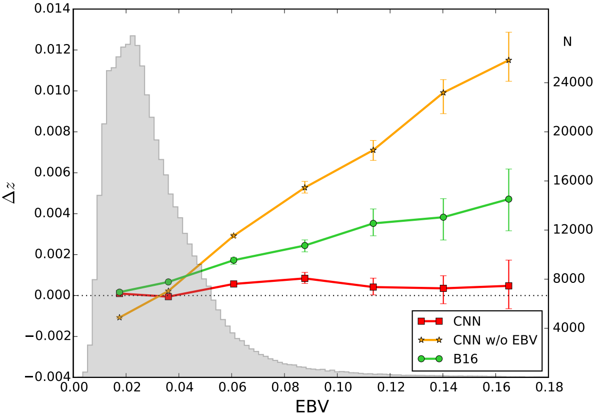
The PIT statistic (Dawid 1984) is based on the histogram of the cumulative probabilities at the true value. For galaxy i with spectroscopic redshift z_i in the test sample :

$$\mathrm{PIT}_i = \int_{-\infty}^{z_i} PDF_i(z)dz$$



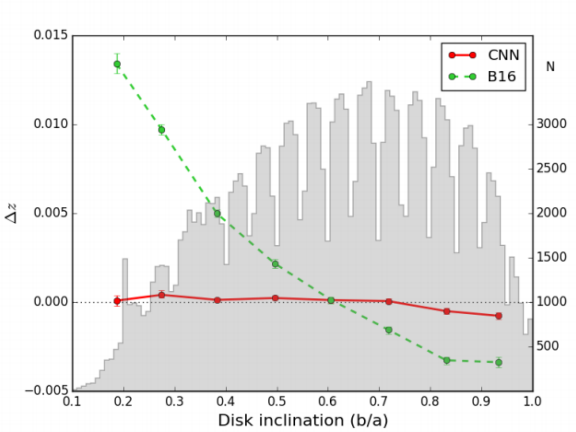
Impact of the extinction of our Galaxy on photometric redshifts

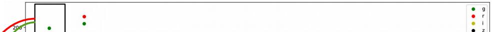
Our method tends to overestimate redshifts in obscured regions (confusing galactic dust attenuation with redshift dimming), unless $E_{(B-V)}$ is used for training



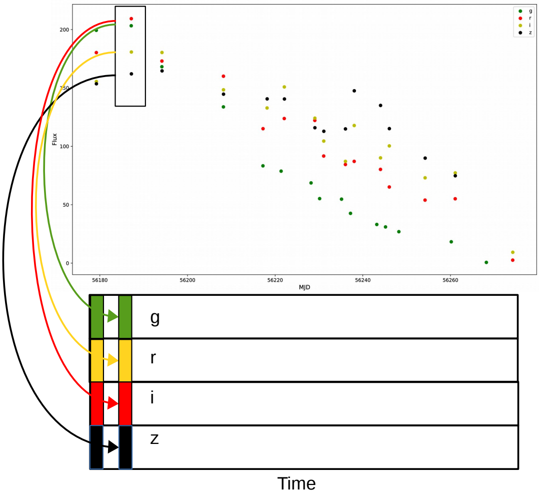
Impact of the disk inclination of galaxies on photometric redshifts

Our method automatically corrects for galactic dust reddening which increases with disk inclination





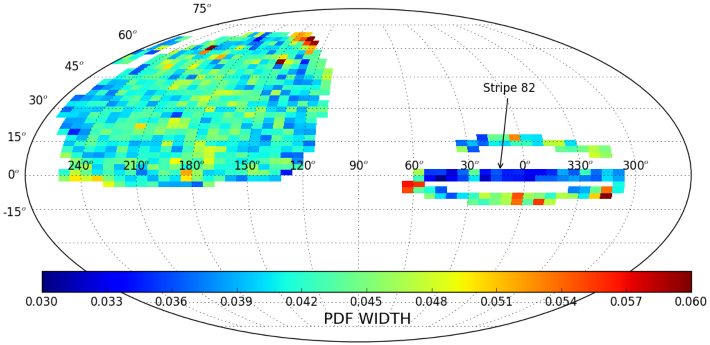
The Light Curve Image (LCI)



⚠ Overfitting of missing data (zero values)

Impact of Signal-to-Noise Ratio (SNR) on widths of PDFs

The Stripe 82 region, which combines repeated observations of the same part of the sky, gives us the opportunity to look into the impact of SNR



Projection of features